2025 International Automatic Control Conference (CACS 2025)
Hsinchu, Taiwan, November 5-8, 2025

Multi-TW: Benchmarking Multimodal Models on Traditional Chinese Question Answering in Taiwan

Jui-Ming Yao, Bing-Cheng Xie, Sheng-Wei Peng, Hao-Yuan Chen, He-Rong Zheng, Bing-Jia Tan, Peter Shaojui Wang, and Shun-Feng Su, *fellow IEEE*

***Abstract*—Multimodal Large Language Models (MLLMs) process visual, acoustic, and textual inputs, overcoming the limitations of single-modality LLMs. However, existing benchmarks often neglect tri-modal evaluation in Traditional Chinese and overlook inference latency. To fill this gap, we introduce Multi-TW, the first Traditional Chinese benchmark for evaluating the performance and latency of any-to-any multimodal models. Multi-TW comprises 900 multiple-choice questions (image & text, audio & text pairs) from authentic proficiency tests developed with the Steering Committee for the Test of Proficiency-Huayu (SC-TOP). We evaluated various any-to-any models and vision-language models (VLMs) with audio transcription. Our findings show closed-source models generally outperform open-source ones across modalities, though open-source models can excel in audio tasks. End-to-end any-to-any pipelines demonstrate significant latency advantages over VLM with separate audio transcription. Multi-TW offers a holistic view of model capabilities, underscoring the need for Traditional Chinese fine-tuning and efficient multimodal architectures.**

I. Introduction

Pre-trained Large Language Models (LLMs), such as LlaMA [7], [8] and Qwen [10], [11], [12], have demonstrated remarkable success across a wide range of natural language processing (NLP) tasks. However, these text-only models remain constrained by their single-modality input. To address this limitation, recent research has increasingly focused on Multimodal Large Language Models (MLLMs), which can jointly process and reason over visual, acoustic, and textual inputs [1], [2].

In the visual domain, models such as CLIP [14] and Flamingo [15] have shown that contrastive pretraining and multimodal fusion architectures enable state-of-the-art zero-shot image classification, image captioning, and few-shot visual reasoning [3], [4]. Building upon these breakthroughs, Vision-Language Models (VLMs) like LLaVA [16] have pushed the frontier further, inspiring fine-tuned successors such as Vicuna [17] and Alpaca [23], which expand multimodal reasoning capabilities across broader task domains. The models evaluated in our experiments, such as the LLaVA series, PaliGemma 2 [22], Idefics2 [21], Llama 3.2-Vision [18], UI-TARS [19] and Qwen VL [20] series, represent the cutting edge in these developments.

With the evolution of VLMs, increasing attention has turned toward audio-language modeling. Audio Language Models (ALMs) typically employ an audio encoder that transforms raw waveform signals into token representations that can be processed by a language model [3], [4]. For instance, Qwen-Audio [26] and Qwen-Audio2 [27] utilize the Qwen model series [10], [11] as their language modeling backbone and incorporate OpenAI's Whisper [24] for end-to-end speech recognition. Other architectures, such as AudioPaLM [28], fuse the text-based capabilities of PaLM-2 [13] with the discrete audio token modeling of AudioLM [25], enabling both high-quality speech recognition and speech-to-speech translation in a unified framework.

More recently, research has progressed toward universal any-to-any multimodal models that support cross-modal input and output across vision, audio, and text. Prominent examples include NExT-GPT [29], AnyGPT [30] and Unified-IO 2 [31], all pushing the limits of unified multimodal intelligence. Later, this trend transferred into multilingual support, as shown in open-source models like Baichuan-Omni-1.5 [32] and Qwen2.5-Omni [33], as well as closed-source systems such as Gemini, which achieve strong performance in both Chinese and English understanding.

To rigorously evaluate the capabilities of such models, several benchmarks have been proposed. However, most evaluations still assess only two modalities at a time. For instance, NExT-GPT [29] and AnyGPT [30] focus on pairwise modality evaluations. Recently, Qwen2.5-Omni [33] and Baichuan-Omni-1.5 [32] have adopted OmniBench [34], a tri-modal benchmark designed to assess performance across text, image, and audio simultaneously, providing deeper insight into a model's unified reasoning ability.

Despite these advances, a critical gap remains in the evaluation of multimodal models in Traditional Chinese. **Existing Traditional Chinese benchmarks are largely text-based.** TMMLU [35] and its extension TMMLU+ [36] provide comprehensive text-only evaluations of LLMs. VisTW [38] moves into the multimodal space by evaluating VLMs on multiple-choice and dialog-based tasks; however, no benchmark currently supports comprehensive evaluation across textual, visual, and acoustic modalities in Traditional Chinese. In addition to this linguistic gap, we observe that most existing benchmarks prioritize accuracy, often overlooking model inference time. This approach is insufficient for real-world applications where both accuracy and efficiency are crucial.

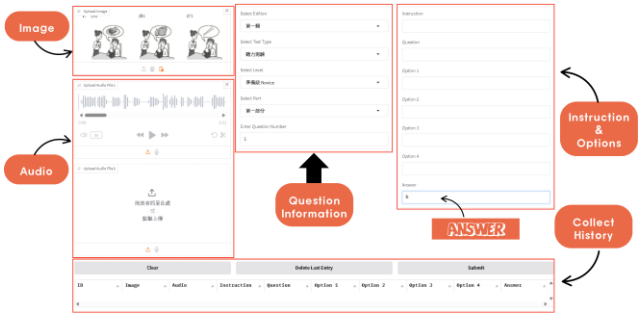

Figure 1. Illustration of data collection interface.

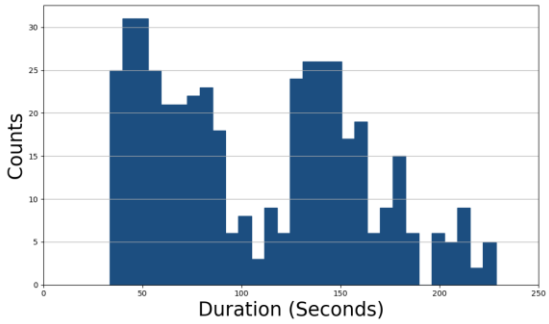

Figure 2. Distribution of audio durations in Multi-TW.

To address this gap, we introduce **Multi-TW**, the first benchmark specifically designed for evaluating the performance and latency of any-to-any multimodal models in Traditional Chinese. Multi-TW consists of image-text and audio-text pairs, enabling rigorous evaluations that cover textual, visual, and acoustic modalities. Datasets are available on Hugging Face:

https://huggingface.co/datasets/ntuai/multi-tw.

In summary, our contributions are as follows:

- We propose **Multi-TW**, the first Traditional Chinese benchmark for rigorous evaluation across text, audio, and visual inputs.
- We collaborated with the Steering Committee for the Test of Proficiency-Huayu to incorporate authentic, real-world assessment tasks into our machine evaluation framework.
- We conducted comprehensive experiments on both any-to-any models and VLMs (the latter using ASR for audio input).
- In addition to accuracy, we evaluate latency to offer a more holistic view of model performance in real-world settings.

II. Multi-TW Benchmark

A. Data Construction

To construct the Multi-TW dataset, we collaborated with the Steering Committee for the Test of Proficiency-Huayu (SC-TOP), a dedicated agency responsible for developing and promoting Taiwan's Mandarin proficiency tests for non-native speakers. These exams, primarily in a multiple-choice format, underwent rigorous utility analysis to ensure their practical value and effectiveness.

The construction phase spanned from September 2023 to December 2023, primarily using publicly available data. All items in Multi-TW underwent a standardized collection and processing workflow performed by our research team to ensure consistency and accuracy. We developed an interface to accelerate data collection and automate labeling, as depicted in Fig.1. Initially, purely textual questions were removed. The remaining items, which involved various combinations of modalities, were then curated to form image-text and audio-text pairs. To address data imbalance and expand the image-text subset, some questions originally coupling image and audio were adapted by extracting their ground-truth audio transcripts, which were then paired with the corresponding image as the textual component. Subsequently, each image-text and audio-text multiple-choice item was serialized into a unified JSON schema, containing the original question, response options, instructions, and references to the separately stored image or audio files.

B. Quality Control

To ensure data integrity, each image-text and audio-text pair was independently reviewed by a second annotator to verify content consistency and accuracy, ensuring the absence of syntax errors, missing information, or incorrect answer choices. Our quality control process involved four stages. First, a completeness check confirmed that each question contained all required components: text (prompt, options, and solution index), an image or audio file, and associated metadata. Entries with missing or inconsistent elements (e.g., a mismatched file name) were flagged and corrected. Second, we validated file consistency. Each image was viewed to confirm it was properly formatted (150 dpi PNG), and each audio clip was played to ensure audible clarity in the specified 128 kbps MP3 (or other, specify format) setting. Invalid or corrupted files were replaced or re-processed. Third, we verified label accuracy by aligning the text content with the corresponding modality. For the image-text subset, the visual context had to align with the question stem and options (e.g., an illustration of a given scenario). For audio-text items, the spoken content was compared with the multiple-choice options to verify that the designated answer was correct. After all corrections were made, each question was subjected to a final review to verify that the files and metadata were correctly updated. Only after passing this final check was the question approved for inclusion in the final dataset.

C. Data Analysis

Multi-TW comprises 900 multiple-choice questions curated to assess Traditional Chinese proficiency in a multimodal context. The dataset is equally divided into 450 image-text items and 450 audio-text items. In the following sections, we refer to these as 'vision-based items' and 'audio-based items,' respectively. This balanced design enables direct comparison of model performance on visual versus auditory modalities paired with Traditional Chinese

TABLE I.
COMPARISON OF MULTI-TW WITH OTHER DATASETS. FOR ALM-BENCH, WE ONLY COMPARE THE SUBSET FOR TRADITIONAL CHINESE. (A: AUDIO, T: TEXT, V: VISION) (TRADITIONAL CHINESE: ZH, ENGLISH: EN)

Dataset	Modalities	Language	Test size	Subjects
TMMLU+	T	zh	20,118	66
ALM-Bench	T, V	zh	52	13
VisTW-MCQ	T, V	zh	3,795	21
OmniBench	A, T, V	en	1,142	8
Multi-TW (Ours)	A, T, V	zh	900	7

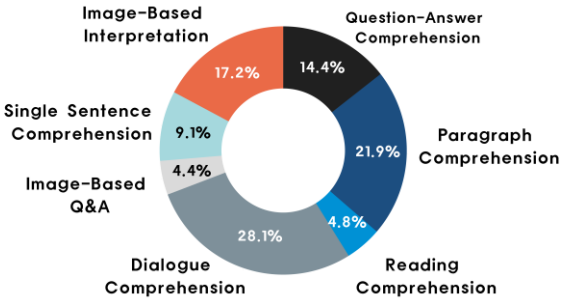

Figure 3. Distribution of question types in Multi-TW.

text and encourages the development of models that handle both input types proficiently.

The vision-based subset features 397 distinct images and includes 407 three-choice items alongside 43 four-choice items. These images depict contextual illustrations, diagrams, and real-world scenarios. All audio-based items employ a four-choice format. Consequently, the 900-item benchmark comprises 407 three-choice questions and 493 four-choice questions (43 from vision-based and 450 from audio-based). For the audio-based items, the average question length is approximately 12 words, and the average option length is approximately 10 words. The average duration of the audio is approximately 107.5 seconds, as illustrated in Fig.2.

D. Comparison with Existing Benchmarks.

Multi-TW evaluates multimodal understanding by measuring performance on two primary task families: vision-based tasks and audio-based tasks. Details of subtask distribution are provided in Fig.3, and representative samples are shown in Fig. 4. This diverse mix of task types ensures that Multi-TW evaluates a broad spectrum of multimodal understanding capabilities.

Table I presents a comparison of Multi-TW with other notable Traditional Chinese language evaluation datasets. While existing benchmarks like TMMLU+ [36] focus on text-only LLM capabilities, and VisTW-MCQ [38] and ALM-Bench [37] incorporate vision and text, Multi-TW, to the best of our knowledge, is the first benchmark to provide comprehensive image-text and audio-text evaluation for Traditional Chinese, thereby covering visual, textual, and auditory modalities. By unifying these input types within a

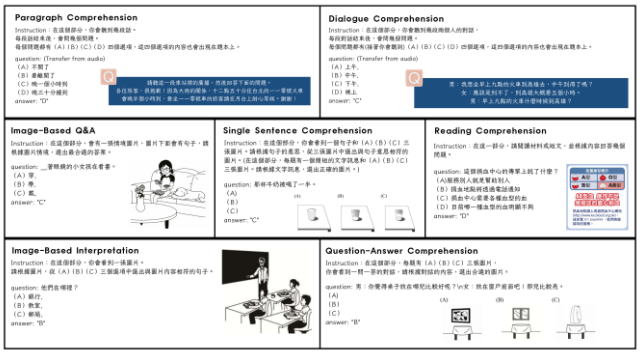

Figure 4. Illustration of samples from the Multi-TW dataset.

A single benchmark framework for Traditional Chinese fills a critical gap and enables a more holistic evaluation of multimodal models. Moreover, beyond its rich modality and linguistic features, Multi-TW's audio samples average 107.5 seconds in length, substantially longer than the 9.12 seconds typical of OmniBench [34] (which primarily tests English). This extended duration enables a more rigorous evaluation of long-form listening comprehension abilities.

III. EXPERIMENTS

To demonstrate the utility of Multi-TW and establish initial performance benchmarks, we conducted experiments using a variety of publicly available multimodal language models. This section details our experimental setup, the models evaluated, and the observed results.

A. Experiment Setup

All experiments were conducted on an NVIDIA A100-SXM4 80GB GPU. All 900 questions in Multi-TW were used for evaluation in a zero-shot setting. The evaluation metric reported is exact-match accuracy, reflecting the percentage of correctly answered multiple-choice questions. We detail our prompting strategy, time measurement protocols, and model selection below.

B. Prompting Strategy.

For all evaluated models, a uniform prompt was appended to each question. This prompt instructs the model to directly output a single character representing the chosen option, without any additional explanation or reasoning. The general prompt template provided to the models is as follows: {question} 僅輸出正確答案的字母，格式必須為 'A', 'B', 'C', 'D'，輸出限制為單個字母，無需解釋。This prompt instructs the model to directly output a single character representing the chosen option, without any additional explanation or reasoning.

C. Time Measurement.

We recorded the elapsed time for four sequential stages: data loading, data preprocessing, model inference, and metric computation. Data preprocessing and model inference account for the majority of runtime and utilize identical code across all open-source models. Therefore, our timing analysis focuses primarily on the combined duration of these two phases for

open-source models. Closed-source models were omitted from this specific latency analysis, as their response times are dominated by external API calls and network transmission, which are not directly comparable. To eliminate variability from differing output lengths, we fixed the model's maximum generation length to one token for all timed experiments.

D. Model Selection

We evaluated several any-to-any models that process text, image, and audio inputs to generate text output in Traditional Chinese, as well as several VLMs where audio input was provided via ASR transcripts. These models, presented in Tables II and III, span both closed- and open-weight categories and were selected based on their state-of-the-art performance, availability, architectural diversity, and varying degrees of exposure to Chinese language data.

For closed-source any-to-any models, we selected gemini-2.0-flash and gemini-1.5-flash from Google. For open-source any-to-any models, we chose the Qwen2.5-Omni series and Baichuan-Omni-1.5, both pretrained primarily on Simplified Chinese. Although Simplified and Traditional Chinese share lexical similarities, they differ substantially in character forms and orthographic conventions. We also incorporated UnifiedIO-2, an encoder-decoder Transformer pretrained from scratch mostly on English data (with a small multilingual fraction from mC4 [39]), making it a useful test for zero-shot cross-script transfer as it has not been specifically fine-tuned for either Chinese variant. For VLMs, we employed Whisper-large [24] to transcribe audio inputs into text for the audio-text tasks. The selected VLMs include Qwen2.5-VL-7B, Qwen2-VL-7B, Llama-3.2-11B-Vision, UI-TARS-1.5-7B, Idefics2-8b, the LLaVA series, and PaliGemma2. This selection reflects the current landscape and provides a broad overview of VLM capabilities on our benchmark.

IV. Result and Analysis

This section offers a summary of performance across all evaluated models on the 900-item Multi-TW benchmark, comparing accuracy on the image-text and audio-text subsets alongside inference latency.

A. Performance on Any-to-Any Models.

Table II illustrates the results for any-to-any models across overall accuracy, image-text subset accuracy, audio-text subset accuracy, and inference time. Key observations include:

1) The Qwen2.5-Omni series and Baichuan-Omni-1.5, despite being primarily pretrained and fine-tuned on Simplified Chinese, achieve competitive accuracy on Traditional Chinese inputs, particularly on audio-text tasks.

2) In contrast, UnifiedIO-2-XL, with limited exposure to Chinese, often failed to produce meaningful answers. Manual inspection of its responses (when constraining output length to 30 tokens) revealed that in 78 cases the model echoed the first option's Chinese description, and in 807 cases it consistently selected option "A."

3) Qwen2.5-Omni-7B exhibited the longest inference time among the open-source any-to-any models, approximately 30.8\% longer than Baichuan-Omni-1.5 (11B parameters). This suggests that parameter count is not the sole determinant of inference speed.

4) The results reveal a significant performance gap between open-source and closed-source models, especially in the image-text domain, highlighting the urgent need for dedicated Traditional Chinese fine-tuning and more robust vision components in open-source any-to-any models.

TABLE II.
Performance of Any-to-Any Multimodal Models on Multi-TW.

Models	Accuracy			Latency
	Overall	*Image-Text*	*Audio-Text*	*Inference Time (s)*
gemini-2.0-flash	0.8900	0.8800	0.9000	-
gemini-1.5-flash	0.8111	0.7644	0.8578	-
Qwen2.5-Omni-7B	0.6534	0.4156	0.8911	744
Baichuan-Omni-1.5	0.6289	0.4822	0.7756	569
Qwen2.5-Omni-3B	0.5878	0.3377	0.8378	712
UnifiedIO-2-XL	0.2589	0.2600	0.2578	467

TABLE III.
Performance of Vision-Language Models (VLMs) with ASR (Whisper-large) on Multi-TW.

Models	Accuracy			Latency
	Overall	*Image-Text*	*Audio-Transcription*	*Inference Time (s)*
Qwen2.5-VL-7B-Instruct	0.8423	0.8267	0.8578	1216
Qwen2-VL-7B-Instruct	0.8033	0.7822	0.8244	1187
UI-TARS-1.5-7B	0.7823	0.7378	0.8267	2131
Llama-3.2-11B-Vision-Instruct	0.5578	0.4711	0.6444	1308
idefics2-8b	0.4167	0.5156	0.3178	1228
llava-v1.6-mistral-7b	0.4100	0.4178	0.4022	1305
llava-v1.6-vicuna-7b	0.3345	0.4022	0.2667	1302
llava-v1.5-7b	0.3211	0.3911	0.2511	1201
paligemma2-10b-pt-896	0.2600	0.2800	0.2400	1727

B. Performance on Vision Language Models (with ASR).

We evaluated a range of VLMs using Whisper-large for audio transcription. Table III reports overall accuracy, image-text accuracy, audio-transcript-text accuracy, and inference time. Key observations are:

1) Qwen2.5-VL-7B-Instruct and UI-TARS-1.5-7B lead among the evaluated VLMs. The competitive results from these models, developed by organizations with a strong focus

on Chinese AI, suggest that extensive pre-training on relevant Chinese-language corpora is a crucial factor for strong performance.

2) In contrast, models like Llama-3.2-11B-Vision-Instruct, despite their large parameter counts or general multimodal capabilities, exhibit notably lower performance, potentially due to less exposure to Traditional Chinese data or specific task alignments.

C. *Performance on Latency.*

Open-source any-to-any models completed inference in a range of 467–744 seconds for the entire 900-item benchmark. In comparison, VLMs coupled with an ASR pipeline (Whisper-large for audio transcription, then VLM for comprehension) required 1,187–2,131 seconds, reflecting the overhead of the two-stage processing for audio-related tasks. In addition, while closed-source models' runtimes are not directly comparable due to API encapsulation, they generally exhibit higher end-to-end latency in practice for batch processing due to network factors, though individual query latency might be low.

V. Conclusion

To address the gap in evaluating Multimodal Large Language Models capable of processing visual, acoustic, and textual inputs, particularly in Traditional Chinese, we introduced **Multi-TW**, the first benchmark of its kind. This dataset provides new insights into current multimodal large language models' abilities, including their performance and latency on Traditional Chinese tasks.

Our evaluation reveals that while closed-source models generally achieve strong performance across both image and audio modalities, open-source alternatives currently tend to perform better on audio-text tasks compared to image-text tasks when using any-to-any architectures. The VLM plus ASR approach can achieve strong results but incurs higher latency for audio tasks. We also found that end-to-end any-to-any models offer notable latency advantages over cascaded VLM plus ASR pipelines for processing audio inputs. Our findings underscore the need for more appropriate architecture designs and targeted fine-tuning data for robust multimodal integration, especially for Traditional Chinese.

In future work, we will examine how cross-lingual transfer capabilities influence the performance of Simplified Chinese-trained models on Traditional Chinese reasoning tasks. We also plan to evaluate latency under more rigorous, parallelized experimental conditions and explore alternative settings, such as streaming inference. Furthermore, expanding Multi-TW to include generative tasks and more complex reasoning scenarios will be a key direction.

Acknowledgment

The authors would like to thank the Steering Committee for the Test of Proficiency-Huayu for agreeing to release the Test of Chinese as a Foreign Language data for this study. The responsibility for errors in fact or judgment is ours. We also extend our gratitude to Professor Yun-Nung Chen from National Taiwan University for her invaluable guidance and support.

References

[1] Z. Liang et al., “A survey of multimodel large language models,” Xi’ an, China: Association for Computing Machinery, 2024, pp. 405–409. doi: https://doi.org/10.1145/3672758.3672824.
[2] S. Yin et al., “A survey on multimodal large language models,” 2024, doi: https://doi.org/10.1093/nsr/nwae403.
[3] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, “A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges,” 2025. https://arxiv.org/abs/2501.02189
[4] A. Islam, M. R. Biswas, Wajdi Zaghouani, Samir Brahim Belhaouari, and Z. Shah, “Pushing boundaries: Exploring zero shot object classification with large multimodal models,” 2023. https://arxiv.org/abs/2401.00127
[5] S. Latif et al., “Sparks of large audio models: A survey and outlook,” 2023. https://arxiv.org/abs/2308.12792
[6] J. Peng, Y. Wang, Y. Xi, X. Li, X. Zhang, and K. Yu, “A survey on speech large language models,” 2025. https://arxiv.org/abs/2410.18908
[7] H. Touvron et al., “LLaMA: Open and efficient foundation language models,” 2023. https://arxiv.org/abs/2302.13971
[8] H. Touvron et al., “Llama 2: Open foundation and fine-tuned chat models,” 2023. https://arxiv.org/abs/2307.09288
[9] A. Grattafiori et al., “The llama 3 herd of models,” 2024. https://arxiv.org/abs/2407.21783
[10] J. Bai et al., “Qwen technical report,” 2023. https://arxiv.org/abs/2309.16609
[11] A. Yang et al., “Qwen2 technical report,” 2024. https://arxiv.org/abs/2407.10671
[12] Qwen et al., “Qwen2.5 technical report,” 2025. https://arxiv.org/abs/2412.15115
[13] R. Anil et al., “PaLM 2 technical report,” 2023. https://arxiv.org/abs/2305.10403
[14] A. Radford et al., “Learning transferable visual models from natural language supervision,” 2021. https://arxiv.org/abs/2103.00020
[15] Jean-Baptiste Alayrac et al., “Flamingo: a visual language model for few-shot learning,” 2022. https://arxiv.org/abs/2204.14198
[16] H. Liu, C. Li, Y. Li, and Y. J. Lee, “Improved baselines with visual instruction tuning,” 2024. https://arxiv.org/abs/2310.03744
[17] L. Zheng et al., “Judging LLM-as-a-Judge with MT-Bench and chatbot arena,” 2023. https://arxiv.org/abs/2306.05685
[18] Meta, “Meta llama 3.2-11B vision instruct,” 2024.
[19] Y. Qin et al., “UI-TARS: Pioneering automated GUI interaction with native agents,” 2025. https://arxiv.org/abs/2501.12326
[20] S. Bai et al., “Qwen2.5-VL technical report,” 2025. https://arxiv.org/abs/2502.13923
[21] H. Laurençon, Léo Tronchon, M. Cord, and V. Sanh, “What matters when building vision-language models?,” 2024. https://arxiv.org/abs/2405.02246
[22] A. Steiner et al., “PaliGemma 2: A family of versatile VLMs for transfer,” 2024. https://arxiv.org/abs/2412.03555
[23] R. Taori et al., “Alpaca: A strong, replicable instruction-following model,” Mar. 2023. https://crfm.stanford.edu/2023/03/13/alpaca.html (accessed Jul. 2023).
[24] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022. https://arxiv.org/abs/2212.04356 (accessed May 2025).
[25] Z. Borsos et al., “AudioLM: a language modeling approach to audio generation,” 2023. https://arxiv.org/abs/2209.03143 (accessed May 2025).
[26] Y. Chu et al., “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” 2023. https://arxiv.org/abs/2311.07919 (accessed May 2025).
[27] Y. Chu et al., “Qwen2-audio technical report,” 2024. https://arxiv.org/abs/2407.10759 (accessed May 2025).
[28] P. K. Rubenstein et al., “AudioPaLM: A large language model that can speak and listen,” 2023. https://arxiv.org/abs/2306.12925 (accessed May 2025).
[29] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, “NExT-GPT: Any-to-any multimodal LLM,” 2024. https://arxiv.org/abs/2309.05519 (accessed May 2025).
[30] J. Zhan et al., “AnyGPT: Unified multimodal LLM with discrete sequence modeling,” 2024. https://arxiv.org/abs/2402.12226 (accessed May 2025).

[31] J. Lu et al., “Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action,” 2023. https://arxiv.org/abs/2312.17172v1
[32] Y. Li et al., “Baichuan-Omni-1.5 technical report,” 2025. https://arxiv.org/abs/2501.15368v1
[33] J. Xu et al., “Qwen2.5-Omni technical report,” 2025. https://arxiv.org/abs/2503.20215v1 (accessed May 2025).
[34] Y. Li et al., “OmniBench: Towards the future of universal omni-language models,” 2025. https://arxiv.org/abs/2409.15272
[35] C.-J. Hsu, C.-L. Liu, F.-T. Liao, P.-C. Hsu, Y.-C. Chen, and D. Shiu, “Advancing the evaluation of traditional chinese language models: Towards a comprehensive benchmark suite,” 2023. https://arxiv.org/abs/2309.08448
[36] Z.-R. Tam et al., “An improved traditional chinese evaluation suite for foundation model,” 2024. https://arxiv.org/abs/2403.01858
[37] Ashmal Vayani et al., “All languages matter: Evaluating LMMs on culturally diverse 100 languages,” 2025. https://arxiv.org/abs/2411.16508
[38] Z. R. Tam, Y.-T. Pai, Y.-W. Lee, and Y.-N. Chen, “VisTW: Benchmarking vision-language models for traditional chinese in taiwan,” 2025. https://arxiv.org/abs/2503.10427v2
[39] L. Xue et al., “mT5: A massively multilingual pre-trained text-to-text transformer,” 2021. https://arxiv.org/abs/2010.11934